

The Residual Information Criterion, Corrected

Chenlei Leng *

February 2, 2008

Abstract

Shi and Tsai (JRSSB, 2002) proposed an interesting residual information criterion (RIC) for model selection in regression. Their RIC was motivated by the principle of minimizing the Kullback-Leibler discrepancy between the residual likelihoods of the true and candidate model. We show, however, under this principle, RIC would always choose the full (saturated) model. The residual likelihood therefore, is not appropriate as a discrepancy measure in defining information criterion. We explain why it is so and provide a corrected residual information criterion as a remedy.

KEY WORDS: *Residual information criterion; Corrected residual information criterion.*

1 Introduction

Given n iid observations from a true model

$$y = X\beta_0 + \varepsilon,$$

where $y = (y_1, \dots, y_n)'$, X is a $n \times p$ design matrix, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)'$ follows a multivariate distribution with mean 0 and variance $\sigma_0^2 W(\theta_0)$, and $\beta_0 \in \mathcal{R}^{p \times 1}$ is an unknown vector to be estimated. Here θ_0 is an $m \times 1$ vector parameterizing the correlation matrix. Finally, we denote $\mathcal{A}_0 = \mathcal{A}(\beta_0) = \{j : \beta_{0j} \neq 0, j = 1, \dots, p\}$ as the nonzero coefficient set and $k_0 = \#\mathcal{A}_0$ as the number of nonzero coefficients. The problem of estimating \mathcal{A}_0 is often referred to as variable selection or model selection.

Variable selection in linear regression is probably one of the most important problems in statistics. See for example the references in Shao (1997). To automate the process of choosing a finite dimensional candidate model out of all possible models, various information criteria have been developed. There are two basic elements in all of these criteria: one element that measures the goodness of fit and the other term which penalizes the complexity of the fitted model, usually taken as a function of the parameters used. Generally speaking, the existing variable selection approaches can be classified into two broad categories. On

*Leng is Assistant Professor, Department of Statistics and Applied Probability, National University of Singapore. Leng's research is supported in part by NUS research grant R-155-050-053-133 (Email: stalc@nus.edu.sg).

one hand, AIC type of criteria, such as AIC (Akaike, 1970) and AICc (Hurvich and Tsai, 1989), seek to minimize the Kullback-Leibler divergence between the true and candidate model. On the other hand, BIC (Schwarz, 1978) type of criteria are used to identify a candidate model to achieve selection consistency. Obviously, these criteria are motivated by different assumptions and different considerations, practically and theoretically. Any particular choice on which one to use probably depends on the context and is subject to criticism, as each has its own merits and shortcomings.

In an important paper, Shi and Tsai (2002) proposed an interesting information criterion termed the residual information criterion (RIC). The authors showed that RIC is motivated by the consideration of minimizing the discrepancy between the residual log-likelihood functions of the true and candidate model. However, surprisingly, the authors arrived at a BIC type of criterion, in marked contrast with some other information criteria, such as AIC, AICc, motivated by the same principle of minimizing Kullback-Leibler discrepancy.

In this paper, we show that the RIC approach is not targeting at minimizing the Kullback-Leibler discrepancy between residual likelihoods. We provide a corrected criterion RIC^* motivated by this principle. However, we show that if the residual likelihoods are used to evaluate the Kullback-Leibler divergence between models, RIC (i.e. RIC^*) would always choose the full model. Therefore, the residual likelihood is not an appropriate loss function to define an information criterion. We provide a simple likelihood based approach to circumvent the problem.

The rest of the paper is organized as follows. Section 2 reviews the RIC method in Shi and Tsai. Since Shi and Tsai's RIC is not approximating the Kullback-Leibler divergence, we provide the RIC^* measure as a correction. However, RIC^* always chooses the full model and the reason is explained. Section 3 presents the correct residual likelihood information criterion, motivated by minimizing the Kullback-Leibler divergence between likelihoods instead of residual likelihoods. Concluding remarks are given in Section 4.

2 The Residual Information Criterion

We review the RIC method in Shi and Tsai (2002) in this section. The model we consider in this article is a special case of that in Shi and Tsai (2002) by assuming the Box-Cox transformation parameter λ is 1. The results in the paper can be easily extended to Box-Cox models following similar arguments in Shi and Tsai.

We start by looking at a candidate (working) model

$$y = X\beta + \varepsilon,$$

such that $\#\mathcal{A}(\beta) = k$. We denote the active covariates in X as $X_{\mathcal{A}}$. Inspired by the residual likelihood method in Harville (1974) or Diggle *et al.* (1994) to obtain unbiased estimator for the error variance, we can write the residual log-likelihood as

$$\begin{aligned} L(\theta', \sigma^2) = & -\frac{1}{2}(n-k)\log(2\pi) + \frac{1}{2}\log|X'_{\mathcal{A}}X_{\mathcal{A}}| - \frac{1}{2}(n-k)\log(\sigma^2) - \frac{1}{2}\log|W| \\ & - \frac{1}{2}\log|X'_{\mathcal{A}}W^{-1}X_{\mathcal{A}}| - \frac{1}{2}y'(W^{-1} - H_{\mathcal{A}})y/\sigma^2, \end{aligned} \quad (1)$$

where $H_{\mathcal{A}} = W^{-1}X'_{\mathcal{A}}(X'_{\mathcal{A}}W^{-1}X_{\mathcal{A}})^{-1}X'_{\mathcal{A}}W^{-1}$ and the dependence of W on θ is suppressed. A useful measure of the distance between the working model and the true model is the Kullback-Leibler divergence

$$d(\theta', \sigma^2) = E_0[-2L(\theta', \sigma^2) + 2L_0(\theta'_0, \sigma_0^2)], \quad (2)$$

where E_0 denotes the expectation under the true model and L_0 denotes the residual log-likelihood of the true model. Clearly, the best model loses the least information, in terms of Kullback-Leibler distance, relative to the truth and is therefore preferred. Such a criterion formulates RIC in an information-theoretical framework. Provided that one can unbiasedly estimate $d(\theta', \sigma^2)$, this criterion provides sound basis for parameter estimation and statistical inference under appropriate conditions.

Since $E_0[2L_0(\theta'_0, \sigma_0^2)]$ is independent of the working model, we just need to evaluate $E_0[-2L(\theta', \sigma^2)]$. In Shi and Tsai (2002), (2) is written as

$$\begin{aligned} d(\theta', \sigma^2) = E_0 \Big[(n-k) \log(\sigma^2) + \log |W| + \log |X'_{\mathcal{A}}W^{-1}X_{\mathcal{A}}| \\ + y'(W^{-1} - H_{\mathcal{A}})y/\sigma^2 \Big] \end{aligned} \quad (3)$$

$$\begin{aligned} = (n-k) \log(\sigma^2) + \log |W| + \log |X'_{\mathcal{A}}W^{-1}X_{\mathcal{A}}| \\ + E_0(X\beta_0 + \varepsilon)'(W^{-1} - H_{\mathcal{A}})(X\beta_0 + \varepsilon)/\sigma^2 \end{aligned} \quad (4)$$

by omitting irrelevant terms. By substituting their estimated values $\hat{\theta}$, $\hat{\sigma}^2$ into (4), we have

$$\begin{aligned} d(\hat{\theta}', \hat{\sigma}^2) = (n-k) \log(\hat{\sigma}^2) + \log |\hat{W}| + \log |X'_{\mathcal{A}}\hat{W}^{-1}X_{\mathcal{A}}| \\ + (X\beta_0)'(\hat{W}^{-1} - \hat{H}_{\mathcal{A}})(X\beta_0)/\hat{\sigma}^2 + \text{tr}\{(\hat{W}^{-1} - \hat{H}_{\mathcal{A}})W_0\}\sigma_0^2/\hat{\sigma}^2. \end{aligned} \quad (5)$$

The above expression involves an unknown quantity σ_0^2 . Following Shi and Tsai, we judge the quality of the candidate model by $E_0\{d(\hat{\theta}', \hat{\sigma}^2)\}$. Now, if we assume $\mathcal{A}_0 \subseteq \mathcal{A}$, an assumption also used in deriving AICc (Hurvich and Tsai, 1989), the third term becomes zero. Furthermore, if we assume $\hat{\theta}$ is consistent for θ_0 , we can estimate W_0 by \hat{W} since $\hat{W} = W_0 + o_p(1)$. Then the fourth term can be approximated as $(n-k)\sigma_0^2/\hat{\sigma}^2$. Since $\mathcal{A} \subseteq \mathcal{A}_0$, $(n-k)\hat{\sigma}^2/\sigma_0^2$ then follows χ^2_{n-k} distribution and therefore

$$E_0[(n-k)\sigma_0^2/\hat{\sigma}^2] = (n-k)^2/(n-k-2).$$

Finally, Shi and Tsai argued that $\log |X'_{\mathcal{A}}\hat{W}^{-1}X_{\mathcal{A}}|$ can be approximated by $k \log(n)$. Putting everything together, they proposed the residual information criterion as follows

$$\text{RIC} = (n-k) \log(\hat{\sigma}^2) + \log |\hat{W}| + k \log(n) - k + \frac{4}{n-k-2}, \quad (6)$$

after removing the constant $n+2$. Asymptotically, the complexity part of RIC is of the order $k \log(n)$. Comparing to $\text{BIC} = n \log(\tilde{\sigma}^2) + k \log(n)$, where $\tilde{\sigma}^2$ is the MLE of σ_0^2 , it is intuitively clear that Shi and Tsai's RIC yields consistent models as BIC does. The complexity penalty of RIC, however, is fundamentally different from that of other familiar information criterion such as AIC and AICc, designed to approximate the Kullback-Leibler

divergence between two models. This observation raises the question on whether RIC rightfully approximates the divergence.

It turns out that Shi and Tsai's derivation motivated by minimizing the Kullback-Leibler distance, is incorrect in at least two important places:

1. In (3), a model dependent term $\log |X'_{\mathcal{A}} X_{\mathcal{A}}|$ is omitted from (1), which causes serious bias in deriving an information criterion. In fact, following Shi and Tsai's arguments, we can approximate $\log |X'_{\mathcal{A}} X_{\mathcal{A}}|$ by $k \log(n)$ and thus, RIC should have been

$$\text{RIC}^* = (n - k) \log(\hat{\sigma}^2) + \log |\hat{W}| - k + \frac{4}{(n - k - 2)}.$$

Note that in this formulation, RIC^* always chooses the full model.

2. Even more severely, the practice of approximating the Kullback-Leibler distance between residual likelihoods for comparing models is totally wrong. To illustrate, suppose that $W = I$. In this simple case, the residual likelihood becomes

$$L(\sigma^2) = -\frac{1}{2}(n - k) \log(\sigma^2) - \frac{1}{2} y' [I - X_{\mathcal{A}} (X'_{\mathcal{A}} X_{\mathcal{A}})^{-1} X_{\mathcal{A}}] y / \sigma^2.$$

We see immediately that $E_0[-2L(\sigma^2)] = (n - k) \log(\sigma^2) + (n - k) \sigma_0^2 / \sigma^2$ whenever $\mathcal{A}_0 \subseteq \mathcal{A}$. Thus, for candidate models that include $X_{\mathcal{A}_0}$ in the covariate set, $E_0[-2L(\sigma^2)]$ is always minimized by $\sigma^2 = \sigma_0^2$ and in this case $E_0[-2L(\sigma^2)] = (n - k)(\log(\sigma_0^2) + 1)$. Therefore, if one knows the exact data generating process, the ideal RIC leads to the full model, as its $E_0[-2L(\sigma^2)]$ is the smallest. This explains why RIC^* always chooses the full model.

Given the above serious flaws in going from deriving unbiased estimator of the Kullback-Leibler divergence to RIC, Shi and Tsai's RIC in (6) seems improperly motivated. Fortunately, Shi and Tsai's derivation can be corrected and we introduce a corrected RIC in the next section.

3 A Corrected Residual Information Criterion

Instead of using the residual likelihood, a justifiable criterion is to use the log-likelihood

$$L(\beta', \theta', \sigma^2) = n \log(\sigma^2) + \log |W| + (y - X\beta)' W^{-1} (y - X\beta)$$

in defining the divergence

$$d(\beta', \theta', \sigma^2) = E_0[-2L(\beta', \theta', \sigma^2) + 2L_0(\beta'_0, \theta'_0, \sigma_0^2)].$$

We can write

$$\begin{aligned} E_0[-2L(\beta', \theta', \sigma^2)] &= E_0[n \log(\sigma^2) + \log |W| + (X\beta_0 + \varepsilon - X\beta)' W^{-1} (X\beta_0 + \varepsilon - X\beta)] \\ &= n \log(\sigma^2) + \log |W| + n \sigma_0^2 / \sigma^2 + (X\beta - X\beta_0)' W^{-1} (X\beta - X\beta_0) \sigma_0^2 / \sigma^2. \end{aligned}$$

We can now replace σ^2 , β and θ by their estimates by using the residual likelihood method. Now, suppose that $\mathcal{A}_0 \subseteq \mathcal{A}$. Following Shi and Tsai again, $E_0 n \sigma_0^2 / \hat{\sigma}^2 \approx n(n-k)/(n-k-2)$. Since $\hat{\beta} - \beta_0$ follows normal distribution $N\{0, \sigma_0^2 (X'_{\mathcal{A}} W^{-1} X_{\mathcal{A}})^{-1}\}$ asymptotically,

$$\frac{1}{k} (X\hat{\beta} - X\beta_0)' W^{-1} (X\hat{\beta} - X\beta_0) \sigma_0^2 / \hat{\sigma}^2$$

is distributed approximately as $F(k, n-k)$. Therefore,

$$E_0 \{ (X\hat{\beta} - X\beta_0)' W^{-1} (X\hat{\beta} - X\beta_0) \sigma_0^2 / \hat{\sigma}^2 \} = \frac{k(n-k)}{n-k-2}.$$

Putting everything together, we have the following corrected residual information criterion, which we shall refer to as RICc,

$$\text{RICc} = n \log(\hat{\sigma}^2) + k + \frac{4(k+1)}{n-k-2},$$

by omitting a constant $n+2$. Note that

$$\text{AIC} = n \log(\tilde{\sigma}^2) + 2k,$$

and

$$\text{AICc} = n \log(\tilde{\sigma}^2) + 2n(k+1)/(n-k-2)$$

where $\tilde{\sigma}^2$ is the MLE of σ_0^2 . We can decompose the first expression of RICc, AIC and AICc as $n \log(\text{RSS}) - n \log(n-k)$, $n \log(\text{RSS}) - n \log(n)$ and $n \log(\text{RSS}) - n \log(n)$ respectively. Thus, the complexity penalties for RICc, AIC, AICc are $-n \log(n-k) + k + 4(k+1)/(n-k-2)$, $-n \log(n) + 2k$ and $-n \log(n) + 2n(k+1)/(n-k-2)$ respectively. It can be seen that RICc has a larger penalty function than AIC and a smaller penalty than AICc when $n \gg k$.

4 Concluding Remarks

In fitting a model to data, one is required to choose a set of candidate models, a fitting procedure and a criterion to compare competing models. A minimal requirement for a reasonable criterion is that the population version of the criterion is uniquely minimized by the set of the parameters which generate the data. The population version of the residual likelihood information criterion is minimized by the full model and thus fails to meet this basic requirement. Therefore, the residual likelihood cannot be used as a discrepancy measure between models. A simple remedy is to use the likelihood based Kullback-Leibler divergence.

Being a legitimate criterion on its own, our arguments show that Shi and Tsai's RIC is not motivated by the right principle. Should one have followed their motivation, RIC (i.e. RIC* by our notation) would have always chosen the full model. However, Shi and Tsai's RIC, though motivated by the wrong principle (using the residual likelihood instead of the likelihood) and ignoring dangerously an important term $\log |X'X|$ in approximation, has good small sample performance in their simulations. Additionally, Shi and Tsai's RIC has been successfully applied to a number of applications, such as normal linear regression, Box-Cox transformation, inverse regression models (Ni *et al.*, 2005) and longitudinal data

analysis (Li *et al.*, 2006). The success may be understood as Shi and Tsai's RIC resembles BIC. Despite the increasing popularity of RIC, Shi and Tsai's RIC remains unmotivated. It remains to find a justification for Shi and Tsai's RIC as a future research topic.

References

- Akaike, H. (1970). Statistical predictor identification. *Annals of Institute of Statistical Mathematics*, 22, 203-217.
- Azari, R., Li, L., and Tsai, C.-L. (2006). Longitudinal data model selection. *Computational Statistics and Data Analysis*, 50, 3053-3066.
- Diggle, P.J., Heagerty, P.J., Liang, K.-Y. and Zeger, S.L. (2002). Analysis of longitudinal data. (2nd edition). Oxford: Oxford University Press.
- Harville, D.A. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, 61, 383-385.
- Hurvich, C. M., and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76, 297-307.
- Ni, L., Cook, R. D., and Tsai, C.-L. (2005). A note on shrinkage sliced inverse regression. *Biometrika*, 92, 242-247.
- Scharwz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
- Shao, J. (1997). An asymptotic theory for linear model selection (with discussion). *Statistica Sinica*, 7, 221-264.
- Shi, P. and Tsai, C.-L. (2002). Regression model selection-a residual likelihood approach. *Journal of the Royal Statistical Society B*, 64, 237-252.